Design-based causal inference in bipartite experiments

Sizhu Lu*, Lei Shi*, Yue Fang, Wenxin Zhang, Peng Ding

* Sizhu Lu and Lei Shi contributed equally to this project

Bipartite Experiment

- Bipartite experiments have gained increasing popularity
- Characteristics:
  - Treatment assigned to **treatment/ intervention units**
  - Outcome measured on **outcome units**
  - **Two sets of units are connected through a bipartite graph**

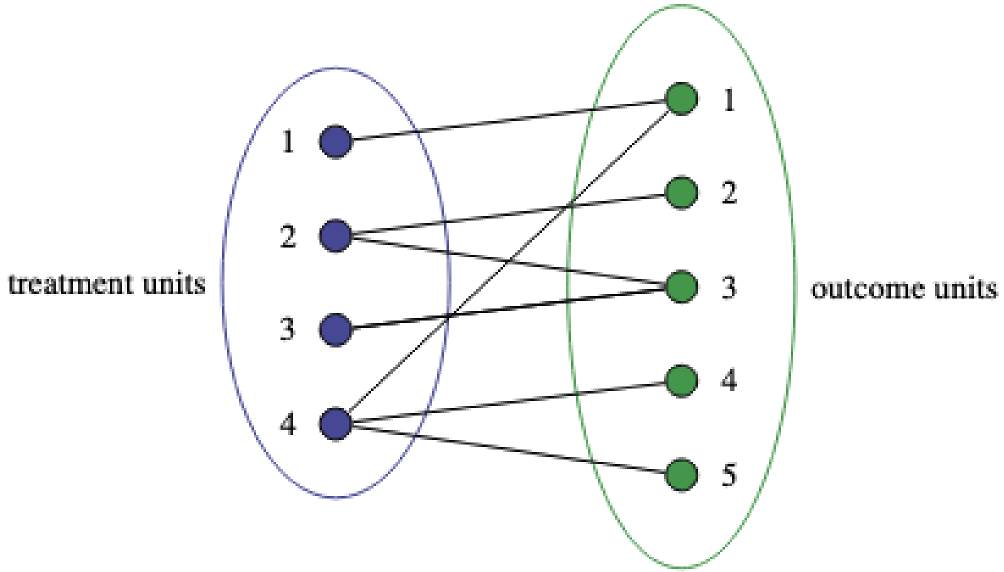
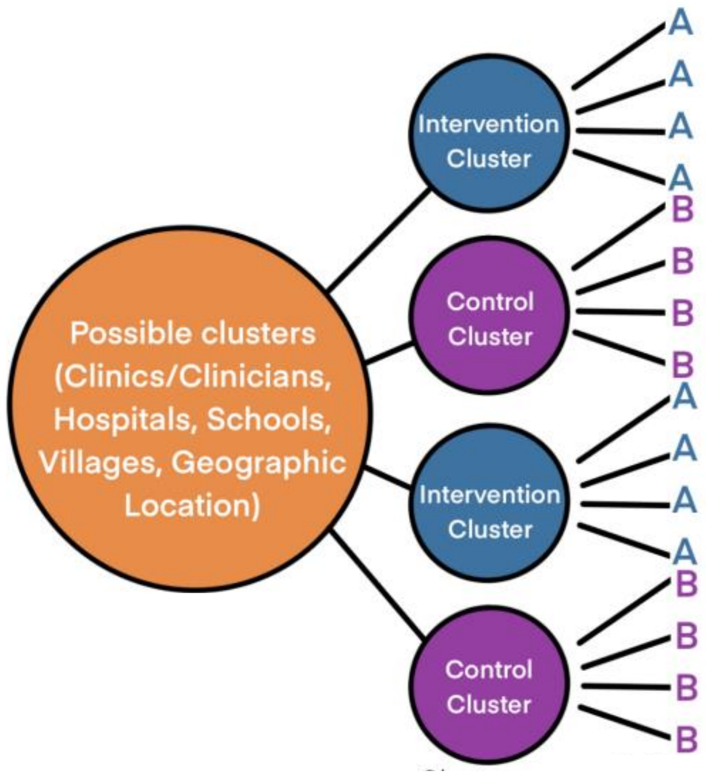


Figure 1: Illustration of a bipartite experiment with $n = 4$ and $m = 5$

Example I: cluster randomization

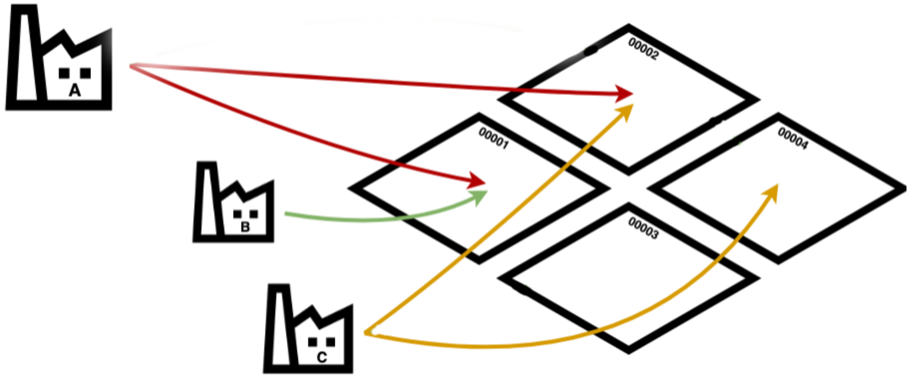
- Experiment setup
  - Units belong to different clusters
  - Treatment units: **clusters**
  - Outcome units: **individuals**
  - Bipartite graph: **cluster membership**
- Example:
  - New digital learning platform in *schools* on *students*' test scores



Example II: hospitalization and power plant

Zigler and Papadogeorgou (2021)

- Experiment setup
  - Selective noncatalytic system positive effect on people's health?
  - Treatment units: **power plants**
  - Outcome units: hospitalization rate at **zip code level**
  - Bipartite graph: **zip codes connect to upwind power plants**



Casual parameter of interest

- Target parameter: total average treatment effect / global average treatment effect

$$\tau = n^{-1} \sum_{i=1}^{n} \{Y_i(\mathbf{1}) - Y_i(\mathbf{0})\}$$

- Widely used in spatial experiments, bipartite experiments, and generally settings with interference
- Of policy interest – all versus nothing comparison
  - All schools use the new platform
  - All power plants launch the new system

Identification challenge and key assumption

- Each unit has 2^m potential outcomes $Y_i(\boldsymbol{z}) = Y_i(z_1, \dots, z_m)$
  - n outcome units, m treatment units
  - Potential outcome framework
  - $Z = (Z_1, \dots, Z_m)$ treatment vector
- Violation of SUTVA: $Y_i(\boldsymbol{Z}) = Y_i(Z_i)$ no longer holds with bipartite interference (not even makes sense anymore)
- Key assumption: generalized SUTVA

Identification challenge and key assumption

- **Generalized SUTVA**: the potential outcomes of unit i depend only on the treatment status of the groups to which it belongs
- Mathematically, $Y_i(\boldsymbol{z}) = Y_i(\boldsymbol{z}_{\mathcal{S}_i})$, where
    - $\boldsymbol{z}_{\mathcal{S}_i}$ is the subvector of treatment for $\mathcal{S}_i$
    - $\mathcal{S}_i$ includes the groups unit i belongs to
- **NO** parametric assumptions on exposure mapping and outcome model!

Preview of results

- Design-based causal inference with bipartite interference
  - No parametric exposure mapping or outcome model
  - Randomness purely from design
  - Identification, weighting estimators, and valid inference
  - Covariate adjustment estimator that improves power

Hájek estimator

- T_i: indicator that all groups are treated; C_i: indicator that all groups are control
- IPW identification formula is feasible by design:
  - Weighted by all-treat or all-control probability
  - Can construct HT or Hájek (our focus)
- IPW weighting formula – motivates a Hájek-type estimator

$$\hat{\tau} \;=\; n^{-1}\sum_{i=1}^{n}\frac{T_iY_i}{p^{|\mathcal{S}_i|}}\Big/ n^{-1}\sum_{i=1}^{n}\frac{T_i}{p^{|\mathcal{S}_i|}} - n^{-1}\sum_{i=1}^{n}\frac{C_iY_i}{(1-p)^{|\mathcal{S}_i|}}\Big/ n^{-1}\sum_{i=1}^{n}\frac{C_i}{(1-p)^{|\mathcal{S}_i|}}$$

Consistency

- Assumptions
  - Generalized SUTVA
  - Bernoulli randomization
  - Bounded potential outcomes and covariates
  - $\bar{S} = O(1)\ \ and\ \bar{D}/n = o(1)$
    - $\bar{\mathrm{S}} = \max |\mathcal{S}_i|$, maximum number of groups each unit belongs to is bounded by a constant
    - $\overline{D}$ denotes the maximum number of units each group contains
    - allowed to be growing but at a slower rate than n
    - $\overline{D}/n$ is the maximum relative size
- Under these four assumptions, $\hat{\tau}$ converges in probability to τ

Consistency

- Assumptions in the power plant example
  - New systems are randomly assigned to power plants
  - Hospitalization rates are bounded
  - Each city is affected by at most 5 nearest upwind power plants within 10km
  - Number of cities each power plant affects is growing slower than n

Asymptotic distribution

- Additional assumption: sparse bipartite graph

Assumption 5 (Sparse bipartite graph). *Define groups j_1 and j_2 are connected if there exists at least one unit belonging to both groups. Assume for any group k, the total number of groups that are connected to k is bounded by an absolute constant B:*

$$\sum_{j \in [m] \setminus \{k\}} \mathbb{1}\{j, k \text{ are connected}\} \leq B, \quad k = 1, \ldots, m.$$

- Power plant example:
  - Two power plants are connected only if there is at least one city within a certain distance of both power plants
  - Geographical network guarantees that power plants far away from each other are not connected
- Examples that are likely to violate the assumption

Asymptotic distribution

- More notation: define three matrices for $i, j = 1, \dots, n$,

$$(\Lambda_1)_{i,j} = p^{-|\mathcal{S}_i \cap \mathcal{S}_j|} - 1, \quad (\Lambda_0)_{i,j} = (1-p)^{-|\mathcal{S}_i \cap \mathcal{S}_j|} - 1, \quad (\Lambda_\tau)_{i,j} = \mathbb{1}\{\mathcal{S}_i \cap \mathcal{S}_j \neq \varnothing\}.$$

- Asymptotic normality:
  - $v_n^{-1/2}(\hat{\tau} - \tau) \quad \to \quad \mathcal{N}(0,1)$ in distribution
  - The asymptotic variance

$$v_n \quad = \quad n^{-2}\left\{\boxed{\tilde{\boldsymbol{Y}}(\mathbf{1})^{\mathrm{T}}\Lambda_1\tilde{\boldsymbol{Y}}(\mathbf{1})} + \boxed{\tilde{\boldsymbol{Y}}(\mathbf{0})^{\mathrm{T}}\Lambda_0\tilde{\boldsymbol{Y}}(\mathbf{0})} + \boxed{2\tilde{\boldsymbol{Y}}(\mathbf{1})^{\mathrm{T}}\Lambda_\tau\tilde{\boldsymbol{Y}}(\mathbf{0})}\right\}$$

var from treated + var from control + covariance

Special case I: classic Bernoulli randomization

Example 4 (Classic Bernoulli randomized experiment). *In classic Bernoulli randomization where the randomization units are identical to the outcome units,*

$$\mathcal{S}_i \cap \mathcal{S}_j = \begin{cases} 1, & \textit{if } i = j, \\ 0, & \textit{if } i \neq j. \end{cases}$$

Thus the asymptotic variance in equation (2) *reduces to*

$$v_n \quad = \quad n^{-2} p(1-p) \sum_{i=1}^{n} \left\{ \frac{\tilde{Y}_i(1)}{p} - \frac{\tilde{Y}_i(0)}{1-p} \right\}^2,$$

which recovers the classic result of Bernoulli randomization in Miratrix et al. (2012, Theorem 1).

Special case II: cluster randomization

Example 5 (Cluster randomization). *In a cluster randomization setting with m clusters and the treatment assignment $Z_k \overset{iid}{\sim} \mathrm{Bern}(p)$ for $k = 1, \ldots, m$, we have*

$$\mathcal{S}_i \cap \mathcal{S}_j = \begin{cases} 1, & \text{if } i, j \text{ belong to the same group}, \\ 0, & \text{otherwise.} \end{cases}$$

If we order the units according to the cluster they belong to, then

$$\Lambda_\tau = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{n_m} \end{pmatrix}, \quad \Lambda_1 = \frac{1-p}{p}\Lambda_\tau, \quad \Lambda_0 = \frac{p}{1-p}\Lambda_\tau,$$

where $\mathbf{1}_{n_k}$ is an $n_k \times n_k$-dimensional matrix with all entries equal to 1 and n_k is the total number of units in cluster k for $k = 1, \ldots, m$. Therefore, the asymptotic variance in equation (2)reduces to

$$v_n = n^{-2} p(1-p) \sum_{k=1}^{m} \left[\sum_{i \in \mathcal{D}_k} \left\{ \frac{\tilde{Y}_i(1)}{p} - \frac{\tilde{Y}_i(0)}{1-p} \right\} \right]^2 .$$

Variance estimation and inference

- A variance estimator:

$$\hat{v} \;=\; \left[\left\{ n^{-2} \sum_{i,j} \frac{T_i T_j (Y_i - \hat{\mu}_1)(Y_j - \hat{\mu}_1)(\Lambda_1)_{i,j}}{p^{|S_i \cup S_j|}} \right\}^{1/2} + \left\{ n^{-2} \sum_{i,j} \frac{C_i C_j (Y_i - \hat{\mu}_0)(Y_j - \hat{\mu}_0)(\Lambda_0)_{i,j}}{(1-p)^{|S_i \cup S_j|}} \right\}^{1/2} \right]^2$$

Estimating var from all-treat POs

Estimating var from all-control POs

- Covariance not estimable: counterfactual unobserved
- Not consistent but conservative
- Asymptotically valid for inference!

When consistent variance estimator?

- Consistent variance estimator if and only if

$$\tilde{\boldsymbol{Y}}(\mathbf{1})^{\mathrm{T}}\Lambda_\tau\tilde{\boldsymbol{Y}}(\mathbf{0}) \quad = \quad \{\tilde{\boldsymbol{Y}}(\mathbf{1})^{\mathrm{T}}\Lambda_1\tilde{\boldsymbol{Y}}(\mathbf{1})\}^{1/2}\{\tilde{\boldsymbol{Y}}(\mathbf{0})^{\mathrm{T}}\Lambda_0\tilde{\boldsymbol{Y}}(\mathbf{0})\}^{1/2}.$$

  - Based on the Cauchy--Schwarz inequality
  - Depends on the network and potential outcomes
- Classic Bernoulli randomization

$$\tilde{Y}_i(1) = \zeta_1\tilde{Y}_i(0) \ \mathit{for\ any}\ i = 1,\ldots,n \ \mathit{and}\ \zeta_1 > 0$$

  - Special case: constant treatment effect
- Cluster randomization

$$\textstyle\sum_{i\in\mathcal{D}_k}\tilde{Y}_i(1) = \zeta_2\sum_{i\in\mathcal{D}_k}\tilde{Y}_i(0) \ \mathit{for\ any}\ k = 1,\ldots,m \ \mathit{and}\ \zeta_2 > 0$$

  - Special case: constant cluster-specific treatment effect

Covariate adjustment

- Outcome-unit-level pretreatment covariates X_i are usually available
  - Centered covariates: $\tilde{X}_i$
- Consider linear adjustment:

$$\hat{\tau}(\beta_1, \beta_0) = n^{-1} \sum_{i=1}^{n} \frac{T_i \boxed{(Y_i - \beta_1^{\mathrm{T}} \tilde{X}_i)}}{p^{|\mathcal{S}_i|}} \Big/ n^{-1} \sum_{i=1}^{n} \frac{T_i}{p^{|\mathcal{S}_i|}} - n^{-1} \sum_{i=1}^{n} \frac{C_i \boxed{(Y_i - \beta_0^{\mathrm{T}} \tilde{X}_i)}}{(1-p)^{|\mathcal{S}_i|}} \Big/ n^{-1} \sum_{i=1}^{n} \frac{C_i}{(1-p)^{|\mathcal{S}_i|}};$$

- Motivated by Lin's estimator in complete randomized experiments
- How to choose the proper coefficients?

Covariate adjustment

- Constructing ”pseudo” potential outcomes with the linear adjustment
- The following CLT holds:

$$\boxed{v_n(\beta_1,\beta_0)^{-1/2}\left\{\hat{\tau}(\beta_1,\beta_0)-\tau(\beta_1,\beta_0)\right\}\to\mathcal{N}(0,1)}$$

- The asymptotic variance is given by

$$\boxed{\begin{aligned} v_n(\beta_1,\beta_0) \quad = \quad & n^{-2}\Big[\{\tilde{\boldsymbol{Y}}(\mathbf{1})-\tilde{\boldsymbol{X}}\beta_1\}^{\mathrm{T}}\Lambda_1\{\tilde{\boldsymbol{Y}}(\mathbf{1})-\tilde{\boldsymbol{X}}\beta_1\}+\{\tilde{\boldsymbol{Y}}(\mathbf{0})-\tilde{\boldsymbol{X}}\beta_0\}^{\mathrm{T}}\Lambda_0\{\tilde{\boldsymbol{Y}}(\mathbf{0})-\tilde{\boldsymbol{X}}\beta_0\} \\ & +2\{\tilde{\boldsymbol{Y}}(\mathbf{1})-\tilde{\boldsymbol{X}}\beta_1\}^{\mathrm{T}}\Lambda_\tau\{\tilde{\boldsymbol{Y}}(\mathbf{0})-\tilde{\boldsymbol{X}}\beta_0\}\Big].\end{aligned}}$$

- Is also upper bounded by

$$\boxed{v_{n,\mathrm{UB}}(\beta_1,\beta_0)=\left(\left[n^{-2}\{\tilde{\boldsymbol{Y}}(\mathbf{1})-\tilde{\boldsymbol{X}}\beta_1\}^{\mathrm{T}}\Lambda_1\{\tilde{\boldsymbol{Y}}(\mathbf{1})-\tilde{\boldsymbol{X}}\beta_1\}\right]^{1/2}+\left[n^{-2}\{\tilde{\boldsymbol{Y}}(\mathbf{0})-\tilde{\boldsymbol{X}}\beta_0\}^{\mathrm{T}}\Lambda_0\{\tilde{\boldsymbol{Y}}(\mathbf{0})-\tilde{\boldsymbol{X}}\beta_0\}\right]^{1/2}\right)^2}$$

- Recall formulas with zero coefficients!

Covariate adjustment

- Try to reduce variance of the estimator by choosing the proper β's
- **Key insight: although asymptotic variances of the estimators are not estimable, the differences are!**
- We have: $L(\beta_1, \beta_0) = v_n(\beta_1, \beta_0) - v_n(0,0)$

$$\boxed{L(\beta_1, \beta_0) = n^{-2} \begin{pmatrix} \beta_1 \\ \beta_0 \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \tilde{\boldsymbol{X}}^{\mathrm{T}} \Lambda_1 \tilde{\boldsymbol{X}} & \tilde{\boldsymbol{X}}^{\mathrm{T}} \Lambda_\tau \tilde{\boldsymbol{X}} \\ \tilde{\boldsymbol{X}}^{\mathrm{T}} \Lambda_\tau \tilde{\boldsymbol{X}} & \tilde{\boldsymbol{X}}^{\mathrm{T}} \Lambda_0 \tilde{\boldsymbol{X}} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_0 \end{pmatrix} - 2n^{-2} \begin{pmatrix} \tilde{\boldsymbol{X}}^{\mathrm{T}} \Lambda_1 \tilde{\boldsymbol{Y}}(\mathbf{1}) + \tilde{\boldsymbol{X}}^{\mathrm{T}} \Lambda_\tau \tilde{\boldsymbol{Y}}(\mathbf{0}) \\ \tilde{\boldsymbol{X}}^{\mathrm{T}} \Lambda_0 \tilde{\boldsymbol{Y}}(\mathbf{0}) + \tilde{\boldsymbol{X}}^{\mathrm{T}} \Lambda_\tau \tilde{\boldsymbol{Y}}(\mathbf{1}) \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \beta_1 \\ \beta_0 \end{pmatrix}}$$

- Minimizing the difference leads to most reduction of variance
- Closed form solution for coefficients:

$$\boxed{\begin{pmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_0 \end{pmatrix} = \begin{pmatrix} \tilde{\boldsymbol{X}}^{\mathrm{T}} \Lambda_1 \tilde{\boldsymbol{X}} & \tilde{\boldsymbol{X}}^{\mathrm{T}} \Lambda_\tau \tilde{\boldsymbol{X}} \\ \tilde{\boldsymbol{X}}^{\mathrm{T}} \Lambda_\tau \tilde{\boldsymbol{X}} & \tilde{\boldsymbol{X}}^{\mathrm{T}} \Lambda_0 \tilde{\boldsymbol{X}} \end{pmatrix}^{-1} \begin{pmatrix} \tilde{\boldsymbol{X}}^{\mathrm{T}} \Lambda_1 \tilde{\boldsymbol{Y}}(\mathbf{1}) + \tilde{\boldsymbol{X}}^{\mathrm{T}} \Lambda_\tau \tilde{\boldsymbol{Y}}(\mathbf{0}) \\ \tilde{\boldsymbol{X}}^{\mathrm{T}} \Lambda_0 \tilde{\boldsymbol{Y}}(\mathbf{0}) + \tilde{\boldsymbol{X}}^{\mathrm{T}} \Lambda_\tau \tilde{\boldsymbol{Y}}(\mathbf{1}) \end{pmatrix}}$$

Covariate adjustment

- How to build an estimator from the observed sample?

$$\begin{pmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_0 \end{pmatrix} = \boxed{\begin{pmatrix} \tilde{\boldsymbol{X}}^{\mathrm{T}}\Lambda_1\tilde{\boldsymbol{X}} & \tilde{\boldsymbol{X}}^{\mathrm{T}}\Lambda_\tau\tilde{\boldsymbol{X}} \\ \tilde{\boldsymbol{X}}^{\mathrm{T}}\Lambda_\tau\tilde{\boldsymbol{X}} & \tilde{\boldsymbol{X}}^{\mathrm{T}}\Lambda_0\tilde{\boldsymbol{X}} \end{pmatrix}}^{-1} \boxed{\begin{pmatrix} \tilde{\boldsymbol{X}}^{\mathrm{T}}\Lambda_1\tilde{\boldsymbol{Y}}(\mathbf{1}) + \tilde{\boldsymbol{X}}^{\mathrm{T}}\Lambda_\tau\tilde{\boldsymbol{Y}}(\mathbf{0}) \\ \tilde{\boldsymbol{X}}^{\mathrm{T}}\Lambda_0\tilde{\boldsymbol{Y}}(\mathbf{0}) + \tilde{\boldsymbol{X}}^{\mathrm{T}}\Lambda_\tau\tilde{\boldsymbol{Y}}(\mathbf{1}) \end{pmatrix}}$$

  - The X-X part: no need to estimate as all are observed
  - The X-Y part: use treated sample to plug in for $\tilde{Y}(1)$ and control sample for $\tilde{Y}(0)$
  - Final estimator: plug in the following estimated coefficients

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_0 \end{pmatrix} = \begin{pmatrix} \tilde{X}^{\mathrm{T}}\Lambda_1\tilde{X} & \tilde{X}^{\mathrm{T}}\Lambda_\tau\tilde{X} \\ \tilde{X}^{\mathrm{T}}\Lambda_\tau\tilde{X} & \tilde{X}^{\mathrm{T}}\Lambda_0\tilde{X} \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i,j} \frac{T_i T_j \tilde{X}_i (Y_j - \hat{\mu}_1)(\Lambda_1)_{i,j}}{p^{|S_i \cup S_j|}} + \sum_{i,j} \frac{C_i C_j \tilde{X}_i (Y_j - \hat{\mu}_0)(\Lambda_\tau)_{i,j}}{(1-p)^{|S_i \cup S_j|}} \\ \sum_{i,j} \frac{T_i T_j \tilde{X}_i (Y_j - \hat{\mu}_1)(\Lambda_\tau)_{i,j}}{p^{|S_i \cup S_j|}} + \sum_{i,j} \frac{C_i C_j \tilde{X}_i (Y_j - \hat{\mu}_0)(\Lambda_0)_{i,j}}{(1-p)^{|S_i \cup S_j|}} \end{pmatrix}.$$

Covariate adjustment

- Establishes consistency, asymptotic normality and conservativeness of the asymptotic variance
- Guarantees reduced asymptotic variance – improve power
- Reduction of estimated variance is not theoretically established, but showcased in simulation study

Monte Carlo Simulation

Table 2: Finite sample performance of estimators $\hat{\tau}$ and $\hat{\tau}^{\text{adj}}$.

		naive estimator					covariate adjustment				
Regime	τ	$\hat{\tau}$	$\text{SE}(\hat{\tau})$	$\hat{\text{SE}}(\hat{\tau})$	coverage	power	$\hat{\tau}^{\text{adj}}$	$\text{SE}(\hat{\tau}^{\text{adj}})$	$\hat{\text{SE}}(\hat{\tau}^{\text{adj}})$	coverage	power
R1	0.221	0.223	0.059	0.086	99.7%	82.3%	0.223	0.055	0.080	99.5%	89.3%
R2	0.256	0.255	0.062	0.085	98.8%	92.8%	0.254	0.058	0.079	98.8%	96.0%
R3	0.355	0.358	0.085	0.124	99.6%	90.6%	0.358	0.082	0.119	99.5%	93.4%

Note: For each regime of data generating process, we report the true total treatment effect τ, the two point estimators, their standard error $\text{SE}(\cdot)$, standard error estimator $\hat{\text{SE}}(\cdot)$, the coverage rate of the 95% confidence interval constructed using the conservative variance estimator, and their power.

- R1: homogeneous treatment effect
- R2: heterogeneous treatment effect, not depending on degrees
- R3: heterogeneous treatment effect, depending on degrees

Discussion

- We discussed design-based causal inference with bipartite interference
  - No outcome model or parametric exposure mapping
  - Randomness purely comes from design
  - Identification, estimation and inference are possible under conditions
  - Covariate adjustment improves power
- Future directions:
  - More general causal parameters (combination with exposure mapping)
  - Treatment unit-level covariates
  - Model-assisted regression estimators?

Thank you!

Comments and suggestions are appreciated.

Monte Carlo Simulation

- Simulation settings:
  - $p = 0.5,\ n = 5000, m = 1500, \bar{S} = 5$
  - Three regimes for potential outcomes

Table 1: Three regimes of data generating process

Regime	$Y_i(\mathbf{1})$	$Y_i(\mathbf{0})$
R1	$\mathcal{N}(0.25 + \gamma^{\mathrm{T}} X_i, 1)$	$\mathcal{N}(\gamma^{\mathrm{T}} X_i, 1)$
R2	$\mathcal{N}(\alpha_i + \gamma^{\mathrm{T}} X_i, 1)$	$\mathcal{N}(\gamma^{\mathrm{T}} X_i, 1)$
R3	$\mathcal{N}(0.1\lvert\mathcal{S}_i\rvert + 1.1\gamma^{\mathrm{T}} X_i, 1.5)$	$\mathcal{N}(\gamma^{\mathrm{T}} X_i, 1.5)$